DOCUMENT RESUME

ED 208 030 TM 810 725

AUTHOR Lehrer, Richard
TITLE Assessing Developmental Hypotheses with Cross Classified Data: Log Linear Models.
PUB DATE Apr 81
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Jos Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Developmental Stages; Goodness of Fit; *Mathematical Models; Statistical Analysis; *Statistical Significance
IDENTIFIERS *Developmental Contingency Tables; *Log Linear Models; Markov Processes

ABSTRACT

Log linear models are proposed for the analysis of structural relations among multidimensional developmental contingency tables. Model of quasi-independence are suggested for testing specific hypothesized patterns of development. Transitions in developmental categorizations are described by Markov models applied to successive contingency tables. A discussion of the role of Pearson chi square and log likelihood significance tests in model selection is followed by two illustrative data sets. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208030

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Assessing Developmental Hypotheses with Cross Classified Data: Log Linear Models

Richard Lehrer

State University of New York at Albany

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Lehrer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 725

Paper presented at the annual meeting of the American Educational Research Association, April 1981, Los Angeles, California

ABSTRACT

Log linear models are proposed for the analysis of structural relations among multidimensional developmental contingency tables. Models of quasi-independence are suggested for testing specific hypothesized patterns of development. Transitions in developmental categorizations are described by Markov models applied to successive contingency tables. A discussion of the role of Pearson chi square and log likelihood significance tests in model selection is followed by two illustrative data sets.

Log linear models are a recently developed data analysis technique that provide a means for the analysis of structural relations among multi-dimensional developmental contingency tables. Utilizing Piaget's (1970) definition of structure as a set of transformations governing a process of self regulation, hypotheses can be framed concerning synchrony among developmental domains (Bates, Begnigni, Bretherton, Camiori, and Voltera, 1979; Wohlwill, 1973). For example, one could explore patterns of relationship among perspective taking levels and other domains of social cognition such as friendship (Selman, 1980) or among several linguistic and cognitive classifications (Bates et al., 1979). In addition, changes in structural relations over time, as in Bates et al. (1979) discussion of local homology models for examining relations between gestural and linguistic complexes, are also germane to Piaget's concept of structure.

The organization of this investigation is divided into four sections. In the first section, an outline of the log linear model is presented as well as the rationale for tests of significance. The next section discusses developmental hypotheses with respect to developmental contingency tables. Thirdly, log linear approaches to developmental contingency tables are presented. The fourth section follows with two illustrative examples.

Log Linear Models

Log linear models are structural models, describing cross classified data. The complexity of the data is reflected by the number of parameters in the model describing its structure (Bishop, Fienberg, and Holland, 1975). For a two way contingency table of order I (i = 1,2. . .I) by J (j = 1,2. . .J), the logarithm of the expected count in the Mij-th cell is written in

(1) where log M_{ij} represents the natural logarithm of the expected values and the u parameters are analogous to their counterparts in an analysis of variance model (see Bishop, Fienberg, and Holland, 1975; Everitt, 1977; Fienberg, 1977).

$$(1)\quad \log M_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

The analogy to analysis of variance is made obvious by inspection of the maximum likelihood parameter estimates presented in (2), (3), and (4), where $l_{ij} = \log m_{ij}$ for all i and j.

$$(2)\quad u = \sum_i \sum_j l_{ij}/IJ$$

$$(3)\quad u_{1(i)} = (\bar{l}_{i+}/J) - u \text{ where } l_{i+} = \sum_j l_{ij} \text{ and is the mean for each each row, j}$$

$$(4)\quad u_{12(ij)} = l_{ij} - (u + u_{1(i)} + u_{2(j)})$$

The maximum likelihood estimates presented in (2), (3), and (4) are equivalent across a wide spectrum of sampling assumptions, ranging from the assumption of sample size as a random variable (Poisson sampling) to the assumption of fixed marginal configurations of sums or product multinominal sampling (see Birch, 1963). Marginal sums refer to the total sum for any row or column.

Hierarchical Models

Log linear models are often arranged hierarchically to facilitate comparisons among models. Another model that could apply to a two way table is written in (5), where the parameter estimate analogous to interaction is deleted.

$$(5)\quad \log m_{ij} = u + u_{1(i)} + u_{2(j)}$$

Since the degrees of freedom for any model are equivalent to the number of

parameter estimates subtracted from the total number of cells, the model in (1) is said to be "saturated" whereas (5) is "unsaturated" as it has fewer parameters than data cells. In addition, (5) is subsumed by(1), as it contains all the parameter estimates of (1), less one. Thus, (5) and (1) are hierarchically related.

Notation and Hierarchical Arrangement

If hierarchical models are assumed, each parameter estimate can be expressed convienently by a variable number and a set of brackets, following Fienberg (1977), with the entire model represented by the highest terms in the expression. Thus (1) is expressed as /1̄2̄/ designating "main effects" and "interaction" and (5) is expressed as /1̄/ /2̄/, designating "main effects" without "interaction."

Goodness of Fit

The relative adequacy of a model can be tested by summary measures of the goodness of fit. The Pearson chi square statistic, presented in (6), is asymptotically distributed as χ^2 under the null hypothesis of independence. The log likelihood ratio statistic, G^2, written in (7) is also asymptotically distributed as χ^2, with degrees of freedom appropriate for the model under estimation, although the Pearson statistic has been demonstrated to follow the asymptotic χ^2 distribution more closely when the table is sparse (Larntz, 1978).

(6) $\chi^2 = \sum_i (X_i - M_i)^2/M_i$, where X, M represent observed and estimated values, respectively

(7) $G^2 = 2 \sum_i \log (X_i/M_i)$

Goodness of Fit and Nested Hierarchies

The log likelihood ratio statistic, G^2, is minimized by the process of maximum likelihood estimation and can be partitioned additively if models

are arranged hierarchically. Thus, the relative increment in selecting one model rather than another is indexed by the change in the G^2 statistic (Goodman, 1969). For example, if two linear models are nested, wherein model two has only a subset of the parameter estimates (u-terms) contained in the first model, the G^2 statistic for the second model is presented in (8), and represents a partition into a measure of the distance of the parameter estimates

(8) $G^2(2) = G^2 \left[(2)/(1)\right] + G^2(1)$

of model (2) as compared to model (1), as well as the distance of the parameter estimates for the first model (Bishop, Fienberg, and Holland, 1975; Fienberg, (1977). Model fitting, as in quantitative analysis (Joreskog, 1978), is a process of trading simplicity of expression for explanation of all of the observed data. The addition of parameters will inevitably increase the goodness of fit of the model to the observed data, but the incremental increase in fit obtained with additional parameters, especially for nested models, is apt to be more important than accounting for all observed frequencies. The latter, accounting for all observed frequencies, is possible because log linear models are not stochastic; log linear models do not explicitly represent errors of measurement etc.

Quasi-Independence

Although the log linear model allows for a wide variety of structural models, often an investigator has hypotheses concerning the specific cells that should contain the observed counts. By fixing these "hypothesized" cells to have an a priori value of zero, termed structural zeroes rather than sampling zeroes (Bishop et al., 1975), one could then expect the remaining observations to be independent of variable classification. The degrees of freedom appropriate for a test statistic under the model of quasi independence

is presented in (9), where V is the number of degrees of freedom usually associated with the model for the complete case, Z_e is the number of cells with structural zeroes, and Z_p is the number of zero entries in the expected marginal configurations.

(9) $DF = V - Z_e + Z_p$

Developmental Contingency Tables

Developmental processes can be structured in terms of continuity and change in the organization of behavior. Continuity in organization is often framed as a stage, where a stage is assumed to represent a benchmark or prototypical organization of behavior (Feldman, 1980). Change is ascribed to stage transition and occurs in the context of a general model of equilibration (Piaget, 1970). One concern for developmental psychology is whether or not an individual's organization, or structure, is consistent across domains (cognitive, social) or tasks (object permanence, means-end). In addition to hypotheses concerning relative stage synchrony or asynchrony, a second dimension of structure concerns patterns of change in classification over time.

Typically, particular hypotheses about stage synchrony are operationalized as expectations of all counts in an I x J contingency table, where I ($i = 1,2. . .I$) indexes the number of rows and J ($j = 1,2. . .J$) indexes the number of columns. Table 1 represents the general I x J table. As in (6), X_{ij} denotes an observed count and "event" can denote domains, tasks, or times of observations (Hoffman, 1980).

Insert Table 1 here

Developmental Models: Two Way Tables

Wohlwill (1973) has proposed several possible patterns of observed frequency counts. Model I corresponds to an interpretation of Piaget's theory that assumes synchronous mental development for different domains or tasks. For a two way table, one would expect the diagonal entries to contain all of the observed values; off diagonal entries are treated as errors of misclassification. The entries in Table 2, indexed by I, represent the pattern of expected counts for model I.

Insert Table 2 Here

A second model proposed by Wohlwill was the triangular hypothesis of decalage. For example, if event one is a prequisite for a second event (IIa), or if the first event is a preintervention observation and the second, a post-intervention observation (IIb), one could expect observed entries to reflect a predominance of off diagonal cell entries as well as the diagonal entries of Model I. The cell entries for model IIa are below the diagonal and conversely, are above the diagonal for model IIb, as presented in Table 2. Entries that are not in cells designated by the triangular hypothesis are considered to be a product of errors in measurement.

Developmental Models: Multidimensional Tables

Previous efforts have focused on structural analysis of the two way developmental table (Hoffman, 1980; Thomas, 1977). However, restriction to the two way cross classification is potentially misleading, especially for a complex data set. The problem is analogous to the effects of partial correlation in quantitative analysis; two variables may appear to be related due to a common relationship with a third variable. In addition, questions

concerning (a) changes in multidimensional cross classified structure over time or (b) structural relations among several domains, remain unanswered by reliance or series of two way tables. To illustrate, developmental models for a three way table are presented next. Table 3 represents the general three dimensional developmental contingency table. As in the two way case, I (i = 1,2. . .I) indexes rows, J (j = 1,2. . .J), indexes columns, and for the three dimensional table, K (K = 1, 2. . .K), indexes "slices." X_{ijk} indexes the observed counts.

Insert Table 3 Here

Log Linear Models for Developmental Relations

Several models of relation among events are conceivable; three are proposed for illustrative purposes. Model III is a "diagonal" hypothesis of complete synchrony among domains or tasks, corresponding to Model I in the two way array. If Model III is assumed, expected counts should be localized in the cells on the diagonal of the three way array, X_{IJK} for all i = j = k, and assuming, for simplicity, I = J = K. For a three way array, the saturated model, $/\overline{1}\ 2\ \underline{3}/$, is presented in (10).

(10) $$\log M_{ijk} = u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

If model III represents the true state of events, then the application of structural zeroes to the diagonal cells, followed by an appropriate adjustment in the degrees of freedom, as noted in (9), results in a model of quasi-independence among the three events; $/\overline{1}/\ /\overline{2}/\ /\overline{3}/$, which is presented in (11).

(11) $$\log m_{ijk} = u_{1(i)} + u_{2(j)} + u_{3(k)}$$

Model IV

Model IV, "block diagonal" or partial synchrony, might arise in the case where events two and three are synchronous but exhibit decalage with respect to the first event. This model corresponds to a variation of the triangular hypothesis of the two way table. For example, the first event could be a prerequisite for events two and theee, which, in turn, are synchronous. For I = J = K, observed counts should be localized in X_{ijk} cells where i = j = k and i > (j = k). Again, the model of quasi independence /$\overline{1}$/ /$\overline{2}$/ /$\overline{3}$/, could be applied to the three way array, to test the feasibility of model IV.

Model V

Model V is a model for conditional independence among developmental events. For example, if process C is assumed to underlie response patterns A and B, then levels of A and B should be conditionally independent for each fixed value of C. Conditional independence is analogous to the concept of partial correlation in quantitative analysis, in that the partial correlation between two variables may be small if the effects of a third variable are accounted for. For example, if perspective taking is presumed to be the process by which social structure evolves, then one might expect high relationships among measures of perspective taking and structural relations among other domains of social cognition. Rather than focusing on relative synchrony among domains, conditional independence models could represent a test for the relative contribution of a process, e.g., perspective taking, to a set of response relationships. For a three way array, a model of conditional independence for B and C, considered as the second and third variables, respectively, is presented in (12). Note that interaction terms containing /$\overline{1}$2/ have been assumed to be null for (12); thus, (12) is nested within (10).

(12) $\log m_{ijk} = u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(i,k)} + u_{23(jk)}$

If model V describes the data structure adequately, the G^2 statistic should not be significant at the chosen α level.

Transitions in Time: Contingency Tables and Markov Models

In the course of an intervention study, one might assess the stability of change over time. If changes in state can be displayed as contingency tables, varying from time 1 to time T, then the process can be modeled by Markov chains, where each individual is classified at each time. Markov chain models for cross classified data have two components--an initial probability vector of category probabilities and a matrix of transition probabilities, denoting the probability of category stability or category transition across time. The initial probabilities will be regarded as a reference point, so further discussion will concern the matrix of transition probabilities.

Transition probabilities are easily estimated from cross classified data. Table 4 presents a poll of voting preference and is taken from Bishop et al., (1975, p. 259). The four transition probabilities are estimated by (13) and a cogent derivation is presented in Bishop et al. (1975).

(13) $\hat{P}_{ij} = X_{ij}/X_{i+}$

Insert Table 4 Here

Markov Models and Stationarity

The order of a Markov chain refers to the assumptions concerning the underlying process of change. If the state occupied by an individual at time t depends only on his state at time t-1, the process is said to be first

order. Restricting attention to the first order chain, if the transition probabilities are independent of time, then the process is described as stationary, regardless of order. In other words, the individual's responses are stable across the time interval sampled.

Log Linear Models and Markov Chains

If an event has I categories and is measured at T times (t = 0,1. . .T), the stationarity of the transition probabilities is tested by first arranging the data into an I x I x T table, where each of the T transition arrays, of order I x I, is a layer in the I x I x T table. If time is the third variable, the model of conditional independence presented in (14) should fit the observed counts in the I x I x T table. Conditional independence indicates that time and, for example, voting preference, are independent, conditional on initial voting preference. Similarly, one could envision intervention studies where the process is stationary for the control group, but not for the treatment group.

$$(14)\quad \log M_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}$$

Illustrations

Two illustrative examples are provided to highlight the application of log linear models to developmental contingency tables.

Two Dimensional Array

Table 5 contains an I x J table, adapted from Wohlwill (1973, p. 220), of synchronous progression between levels of concept A and levels of concept B. The application of Model I to these data requires fixing the diagonal entries,

---

Insert Table 5 here

---

where i = j, to zero and testing the log linear model of quasi independence /1/ /2/. Model IIa, the triangular hypothesis, involves structural zeroes for all X_{ij} where i > j, as well as for the diagonal entries, i = j. The model of quasi independence is then fitted to the entries in the table. Table 6 presents the G^2 statistic, with accompanying degrees of freedom, for the model of complete independence, followed by models of quasi independence (I, IIa). The model of synchrony appears to fit these data best as evidenced by inspection of Table 6 and the significance of the difference statistic; G^2 (FULL) - G^2(I) = 167.63 with 5 df.

Table 6 insert here

Three Dimensional Array

An I x J x K table of artificial data is presented in Table 7, with

---

Insert Table 7 here

I = J = K = 3. Model III, complete synchrony, is a test of the log linear model of quasi independence, with structural zeroes for the diagonal cells, where i = j = k. The G^2 statistic for Model III is 35.88 with 17 df. In contrast, Model IV, partial synchrony, involves a test of quasi independence with structural zeroes as in Model III, in addition to cells where i (j = k) (X_{211}, X_{311}, X_{322}). The G^2 statistic for this version of Model IV yields a G^2 statistic of 29.54 with 14 df. The difference statistic, G^2 (III) - G^2 (IV), is 6.34 with 3 df, which is not significant at α = .05, suggesting that Model III may be more appropriate for this data set.

Summary and Discussion

The log linear model is appropriate for the analysis of multi-dimensional, developmental contingency tables. Several models of development can be postulated concerning relations among events, changes over time, or some combination thereof. Models of quasi independence allow for point hypothesis testing while utilization of modular hierarchies may help investigators to formulate clearer hypotheses in relation to prior expectations, particularly in relation to the relative contribution, as indexed by the G^2 statistic, of each event to the total gestalt. The intention of this presentation was to establish a script for the analysis of developmental relations, while allowing for variation and extension to new themes by other investigators.

References

Bates, E., Begnigni, L., Bretherton, I., Camioni, L., and Voltera, V. The emergence of symbols. New York: Academic Press, 1979.

Birch, M.W. Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society, Series B, 25, 220-233.

Bishop, Y.M.M., Fienberg, S., and Holland, P. Discrete multivariate analysis theory and practice. Cambridge, Mass.: MIT Press, 1975.

Everitt, B. The analysis of contingency tables. London: Chapman and Hall, 1977.

Feldman, D. Stage and transition in cognitive developmental research: Getting to the next level. The Genetic Epistemologist (Vol. IV), 1, 1980.

Fienberg, S. The analysis of cross-classified categorical data. Cambridge, Mass.: MIT Press, 1977.

Goodman, L.A. On partitioning chi-square and detecting partial association in three-way contingency tables. Journal of the Royal Statistical Society, Series B, 31, 486-498.

Hoffman, R. Assessing Developmental Stages with a Contingency Table: A Triangular Hypothesis. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 1980.

Joreskog, K.G. Structural analysis of covariance and correlation matrices. Psychometrika, 1978, 43, 443-477.

Larntz, K. Small-sample comparisons of exact levels for chi-square goodness-of-fit statistics. Journal of the American Statistical Association, 1978, 73, 253-263.

Piaget, J. Structuralism. New York: Basic Books, 1970.

Selman, R. The growth of interpersonal understanding. New York: Academic Press, 1980.

Thomas, H. Fitting cross-classification table data to models when observations are subject to classification error. Psychometrika, 1977, 42, 199-206.

Wohlwill, J. The study of behavioral development. New York: Academic Press, 1973.

Table 1

EVENT ONE	EVENT TWO: Stage 1	Stage 2	Stage J
Stage 1	X_{11}		X_{1J}
Stage 2			
⋮			
Stage I	X_{I1}		X_{IJ}

Table 2

EVENT ONE	EVENT TWO: Stage 1	Stage 2	Stage J
Stage 1	I	IIb	IIb
Stage 2	IIa	I	IIb
⋮			
Stage I	IIa	IIa	I

Table 3

EVENT TWO

Stage 1 Stage 2 ----------Stage J

Event ONE EVENT THREE

	Stage 1---K	Stage 1---K	Stage 1---K
Stage 1	X_{111}		
Stage 2			
⋮			
Stage I			X_{IJK}

Table 4

Second Poll

First Poll	Same	Different	Totals
Same	X_{11}	X_{12}	X1+
Different	X_{22}	X_{21}	X2+
Totals	X+1	X+2	N

Table 5

Levels of concept B	Levels of concept A				
	1	2	3	4	5
1	18	2	0	0	0
2	2	16	2	0	0
3	0	2	16	2	0
4	0	0	2	16	2
5	0	0	0	2	18

Table 6

Model 1	G^2	Df
Complete Table	183.3	16
I	15.67	11
IIa	10.15	2

Table 7

	EVENT TWO								
	1			2			3		
	EVENT THREE								
Event One	1	2	3	1	2	3	1	2	3
1	21	2	2	4	3	3	1	4	4
2	7	3	3	3	30	4	2	4	5
3	6	2	1	3	7	3	1	4	26